CUNY SPS DATA 621 - CTG5 - HW1

Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh

February 27, 2019

Contents

1 Data exploration

1.1 Possible writeup for Data Exploration

Professionals and gamblers alike are always seeking to optimize their chances of winning, whether it be sports, games, or their bets on them. Major League Baseball is a multibillion dollar industry where individual teams, players, and those who profit off of their success stand to benefit most from such optimization.

In order to determine the best way to infer whether the 162 games in a baseball team's year will result in more wins overall, data from 1871 to 2006 where each set of values represented a season for an unnamed team, totalling 2,276 records. For each team their number of wins in a given year were given with a maximum possible of 162 wins, in addition to that team's base hits, doubles, triples, homeruns, walks, and strikeouts by batters, batters hit by pitches, bases stolen by batters and the number of times they were caught stealing, the number of errors, double plays, walks, hits, and homeruns allowed, and strikeouts by pitchers.

1.1.1 Data before imputing values

	n	min	mean	median	max	sd
TARGET_WINS	2276	0	80.79086	82.0	146	15.75215
BATTING_H	2276	891	1469.26977	1454.0	2554	144.59120
BATTING_2B	2276	69	241.24692	238.0	458	46.80141
BATTING_3B	2276	0	55.25000	47.0	223	27.93856
BATTING_HR	2276	0	99.61204	102.0	264	60.54687
BATTING_BB	2276	0	501.55888	512.0	878	122.67086
BATTING_SO	2174	0	735.60534	750.0	1399	248.52642
BASERUN_SB	2145	0	124.76177	101.0	697	87.79117
BASERUN_CS	1504	0	52.80386	49.0	201	22.95634
PITCHING_H	2276	1137	1779.21046	1518.0	30132	1406.84293
PITCHING_HR	2276	0	105.69859	107.0	343	61.29875
PITCHING_BB	2276	0	553.00791	536.5	3645	166.35736
PITCHING_SO	2174	0	817.73045	813.5	19278	553.08503
FIELDING_E	2276	65	246.48067	159.0	1898	227.77097
FIELDING_DP	1990	52	146.38794	149.0	228	26.22639

1.1.2 Data after imputing values

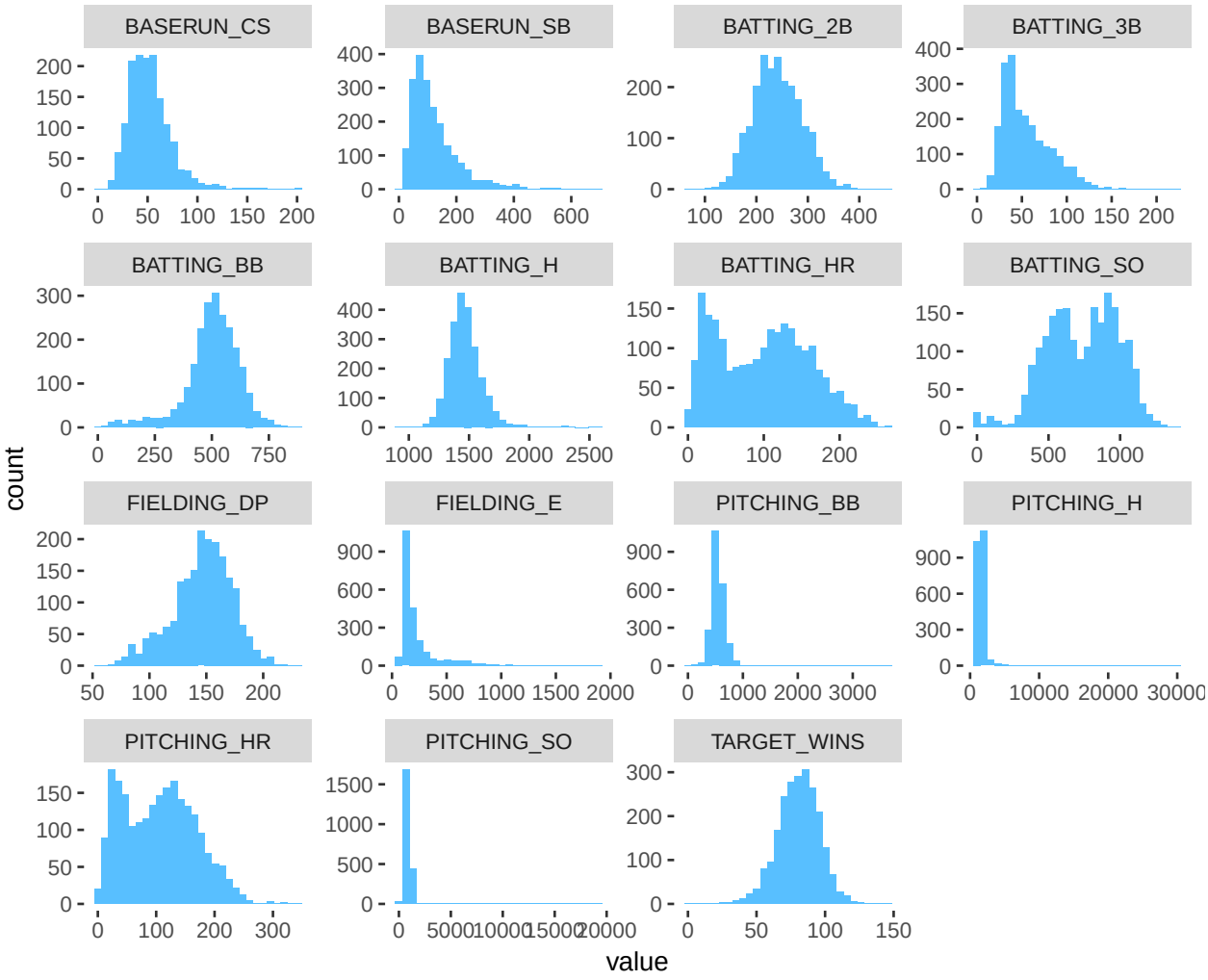
	n	min	mean	median	max	sd
TARGET_WINS	2276	0	80.79086	82.00000	146	15.75215
BATTING_H	2276	891	1469.26977	1454.00000	2554	144.59120
BATTING_2B	2276	69	241.24692	238.00000	458	46.80141
BATTING_3B	2276	0	55.25000	47.00000	223	27.93856
BATTING_HR	2276	0	99.61204	102.00000	264	60.54687
BATTING_BB	2276	0	501.55888	512.00000	878	122.67086
BATTING_SO	2276	0	728.22986	739.00000	1399	246.65040
BASERUN_SB	2276	0	124.62700	105.00000	697	85.28361
BASERUN_CS	2276	0	70.73189	56.42774	201	38.01002
PITCHING_H	2276	1137	1779.21046	1518.00000	30132	1406.84293
PITCHING_HR	2276	0	105.69859	107.00000	343	61.29875
PITCHING_BB	2276	0	553.00791	536.50000	3645	166.35736
PITCHING_SO	2276	0	807.62995	803.50000	19278	543.08096
FIELDING_E	2276	65	246.48067	159.00000	1898	227.77097
FIELDING_DP	2276	52	145.30987	146.00000	228	24.89909

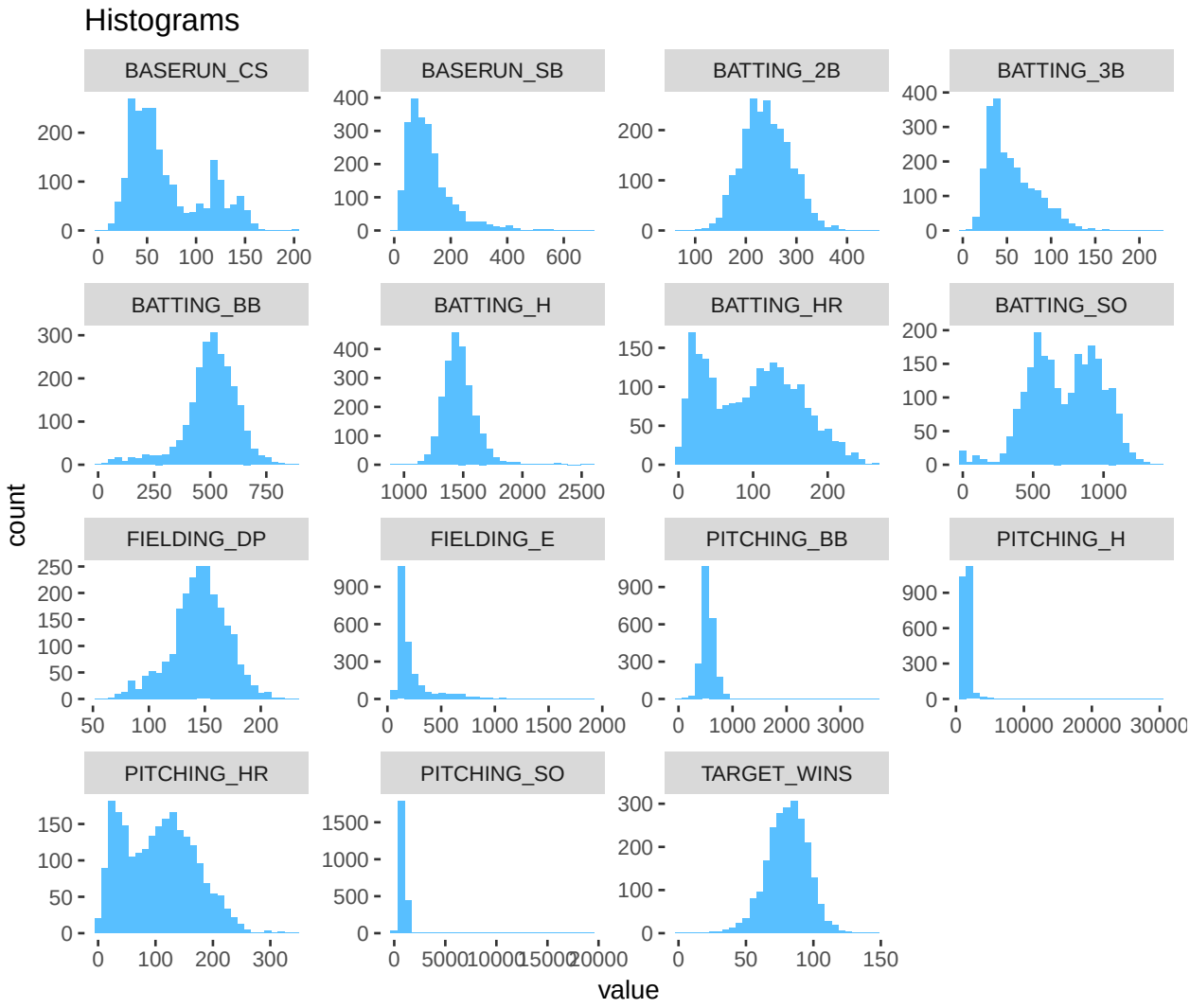
1.1.3 Difference between original and imputed data

	n	min	mean	median	max	sd
TARGET_WINS	0	0	0.0000000	0.000000	0	0.000000
BATTING_H	0	0	0.0000000	0.000000	0	0.000000
BATTING_2B	0	0	0.0000000	0.000000	0	0.000000
BATTING_3B	0	0	0.0000000	0.000000	0	0.000000
BATTING_HR	0	0	0.0000000	0.000000	0	0.000000
BATTING_BB	0	0	0.0000000	0.000000	0	0.000000
BATTING_SO	-102	0	7.3754714	11.000000	0	1.876015
BASERUN_SB	-131	0	0.1347715	-4.000000	0	2.507556
BASERUN_CS	-772	0	-17.9280286	-7.427742	0	-15.053684
PITCHING_H	0	0	0.0000000	0.000000	0	0.000000
PITCHING_HR	0	0	0.0000000	0.000000	0	0.000000
PITCHING_BB	0	0	0.0000000	0.000000	0	0.000000
PITCHING_SO	-102	0	10.1004968	10.000000	0	10.004072
FIELDING_E	0	0	0.0000000	0.000000	0	0.000000
FIELDING_DP	-286	0	1.0780691	3.000000	0	1.327290

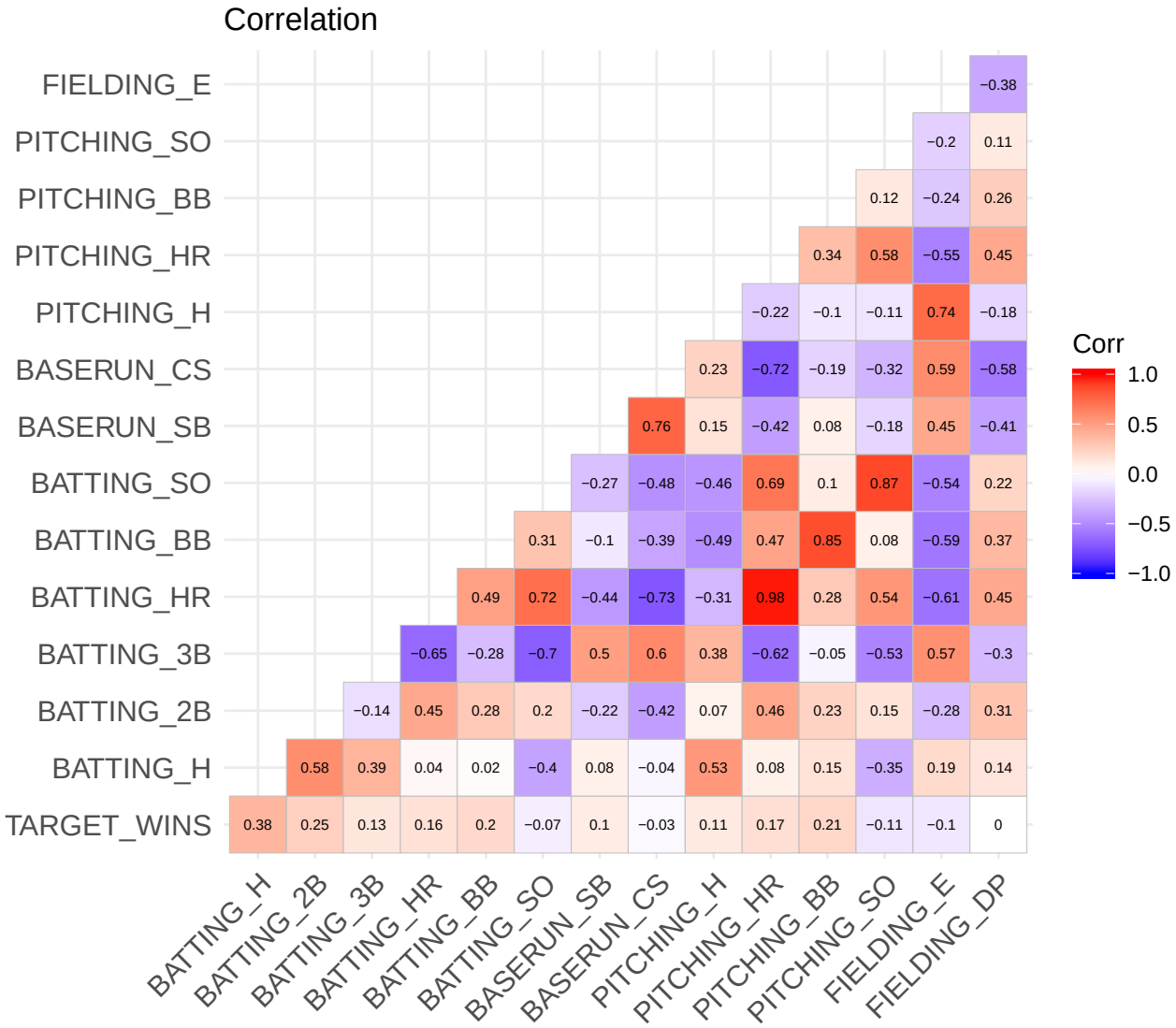
Of all the observations gathered across these fifteen variables, 10.187% were missing information; batters hit by pitches was missing the most, with 2,085 instances of missing information. The standard deviation of the various variables hints at the intense skewing of some of the variables provided, especially the hits allowed, walks allowed and strike outs.

Histograms


BASERUN_CS
BASERUN_SB
BATTING_2B
BATTING_3B
BATTING_BB
BATTING_H
BATTING_HR
BATTING_SO
FIELDING_DP
FIELDING_E
PITCHING_BB
PITCHING_H
PITCHING_HR
PITCHING_SO
TARGET_WINS
count
value



The theoretical effect of strikeouts by batters, batters caught stealing, errors, walks, hits, and homeruns allowed were believed to theoretically have a negative impact on the number of wins of an individual team in a given year. A closer look at the correlation between the variables painted a different picture.



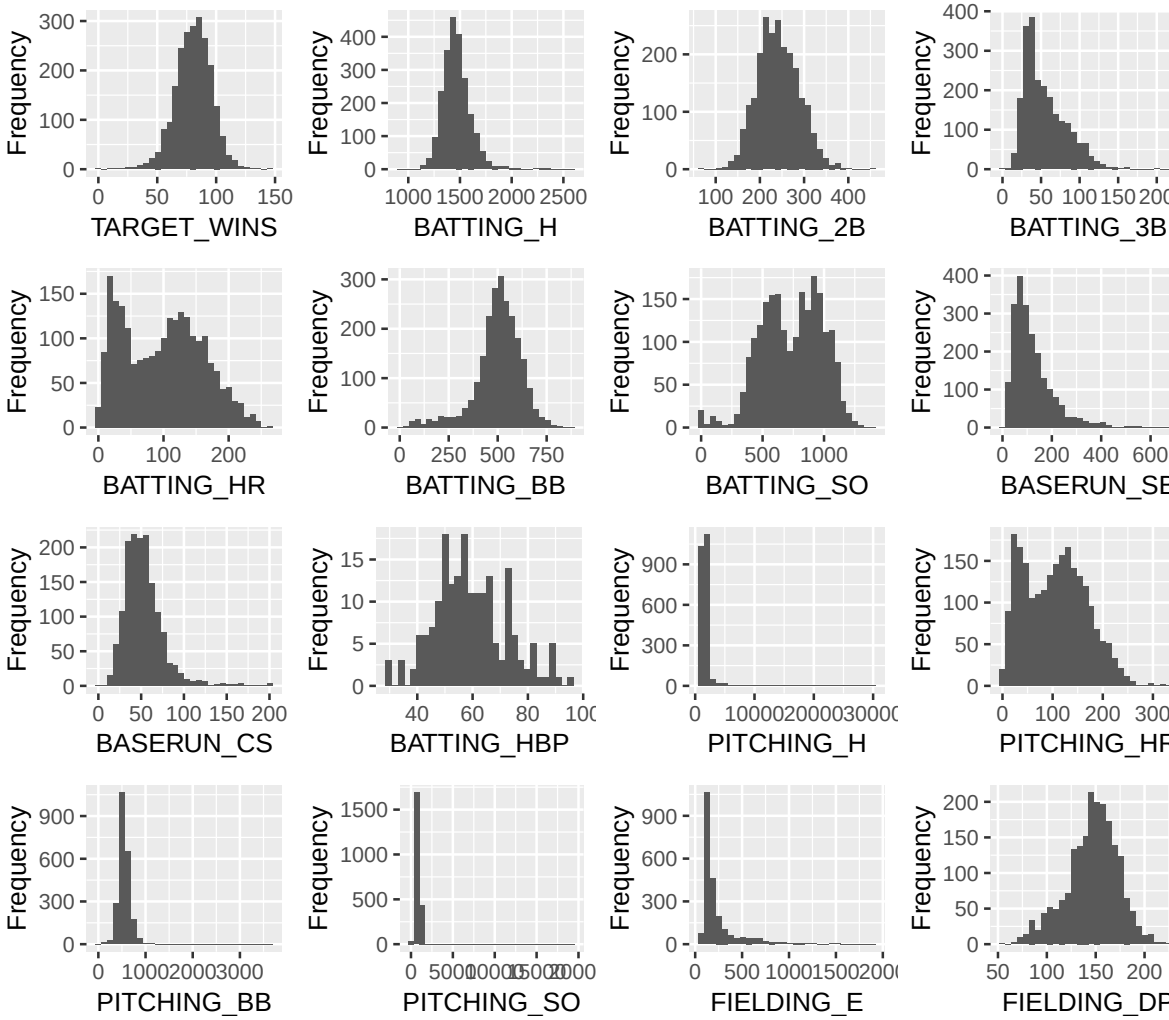
When compared to what was hypothesized, there was actually a positive impact for the number of wins for a team in a given year by walks, hits, and homeruns allowed; at the same time, variables previously thought to have a positive correlation - strikeouts by pitchers and double plays - had a negative correlation for the number of wins. The three variables with the greatest correlation to the number of wins were the hits allowed, the walks by batters, and the walks allowed. Of these, the hits allowed had a relatively low correlation with the walks by batters and the walks allowed, whereas the walks allowed and the walks by batters had a direct positive correlation with one another.

- Describe the size:

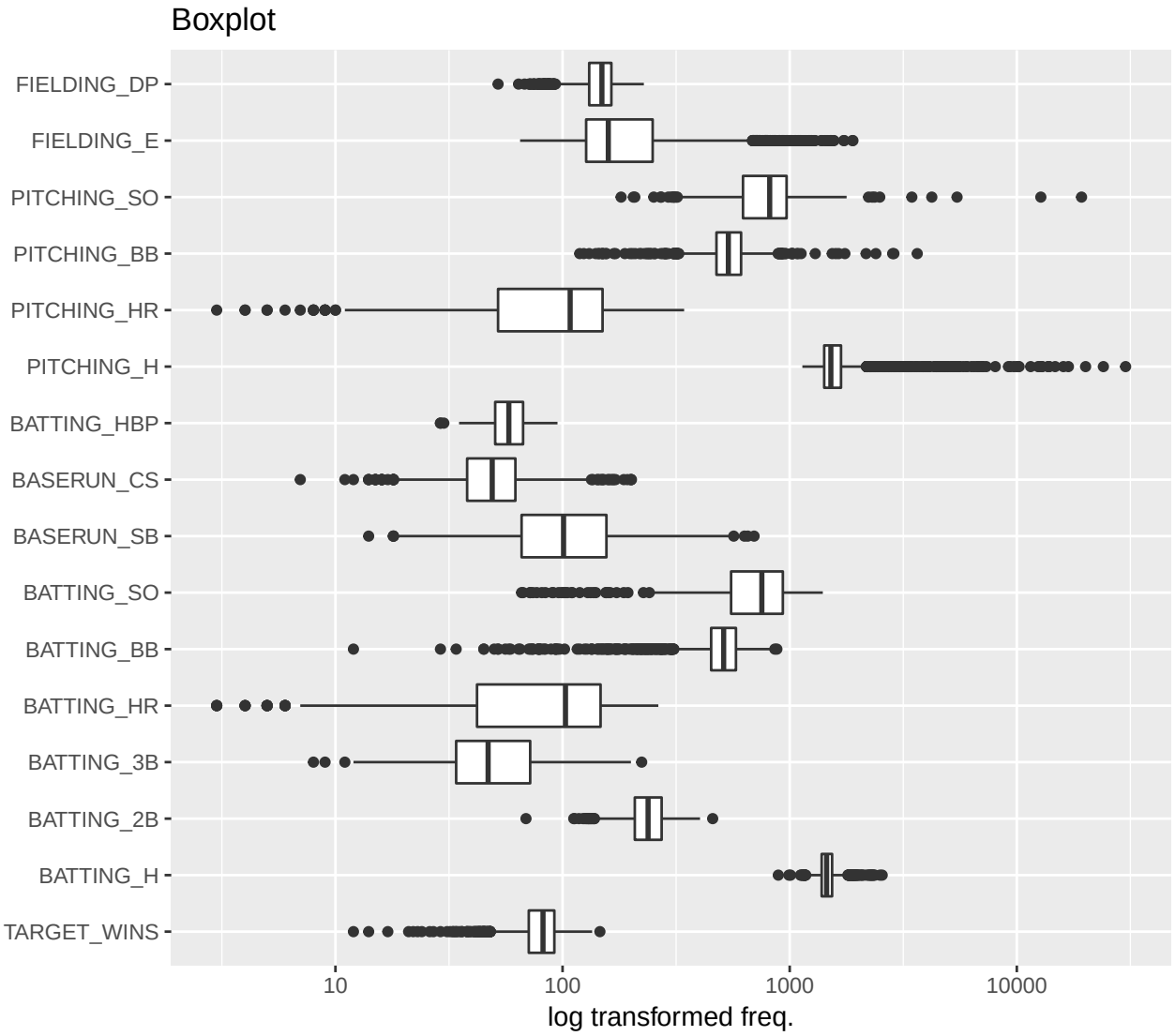
The money ball data is 144kb in size. The data contains 2,276 rows and 16 columns without the index. The variables are continuous integer. The `TARGET_WINS` is our response variable. There are 3,478 missing values out of 36,416 observations.

- Statistics summary

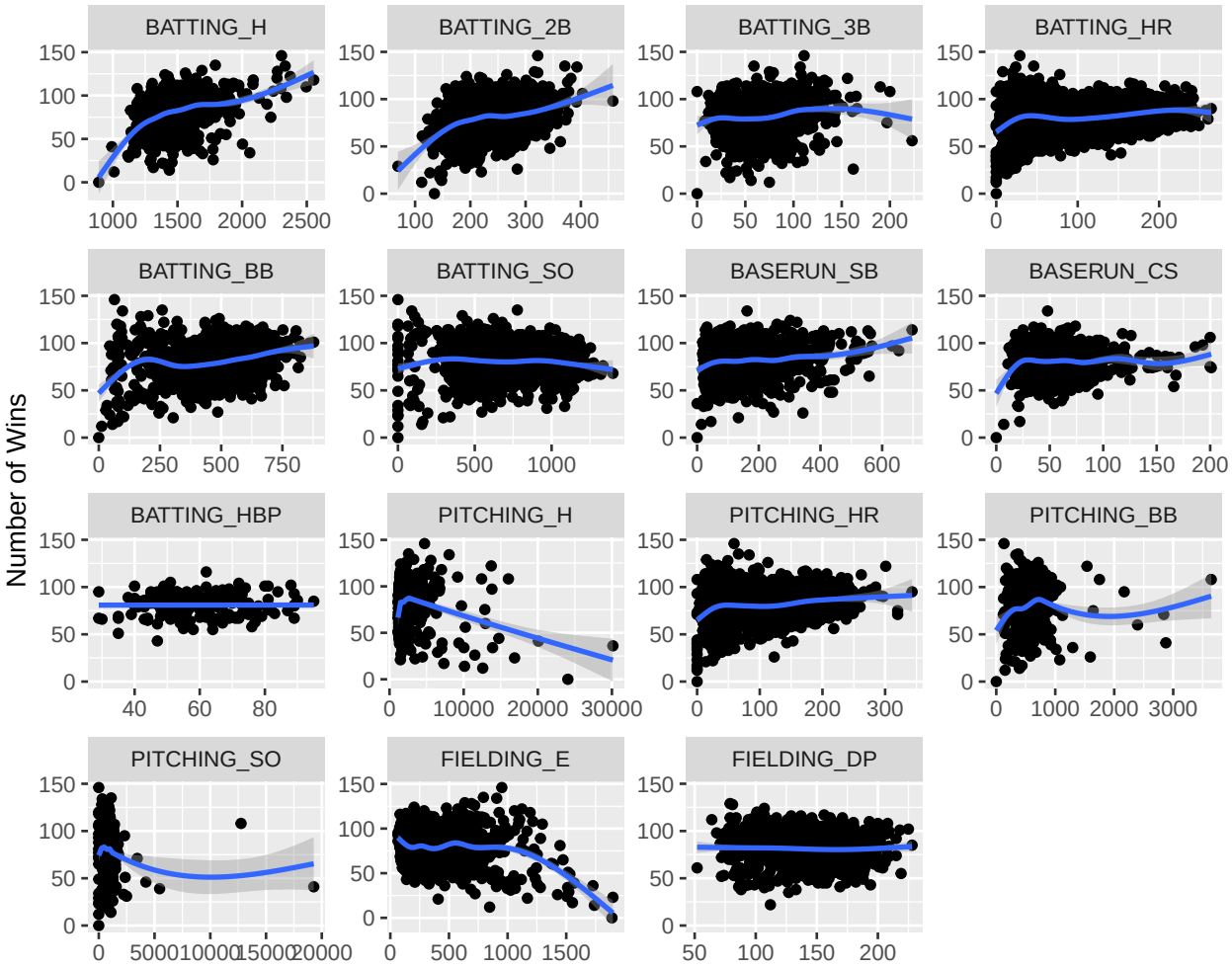
	vars	n	mean	sd	median	trimmed	mad	min	max	range	
TARGET_WINS	1	2276	80.79086	15.75215	82.0	81.31229	14.8260	0	146	146	
BATTING_H	2	2276	1469.26977	144.59120	1454.0	1459.04116	114.1602	891	2554	1663	
BATTING_2B	3	2276	241.24692	46.80141	238.0	240.39627	47.4432	69	458	389	
BATTING_3B	4	2276	55.25000	27.93856	47.0	52.17563	23.7216	0	223	223	
BATTING_HR	5	2276	99.61204	60.54687	102.0	97.38529	78.5778	0	264	264	
BATTING_BB	6	2276	501.55888	122.67086	512.0	512.18331	94.8864	0	878	878	
BATTING_SO	7	2174	735.60534	248.52642	750.0	742.31322	284.6592	0	1399	1399	
BASERUN_SB	8	2145	124.76177	87.79117	101.0	110.81188	60.7866	0	697	697	
BASERUN_CS	9	1504	52.80386	22.95634	49.0	50.35963	17.7912	0	201	201	
BATTING_HBP	10	191	59.35602	12.96712	58.0	58.86275	11.8608	29	95	66	
PITCHING_H	11	2276	1779.21046	1406.84293	1518.0	1555.89517	174.9468	1137	30132	28995	
PITCHING_HR	12	2276	105.69859	61.29875	107.0	103.15697	74.1300	0	343	343	
PITCHING_BB	13	2276	553.00791	166.35736	536.5	542.62459	98.5929	0	3645	3645	
PITCHING_SO	14	2174	817.73045	553.08503	813.5	796.93391	257.2311	0	19278	19278	
FIELDING_E	15	2276	246.48067	227.77097	159.0	193.43798	62.2692	65	1898	1833	
FIELDING_DP	16	1990	146.38794	26.22639	149.0	147.57789	23.7216	52	228	176	

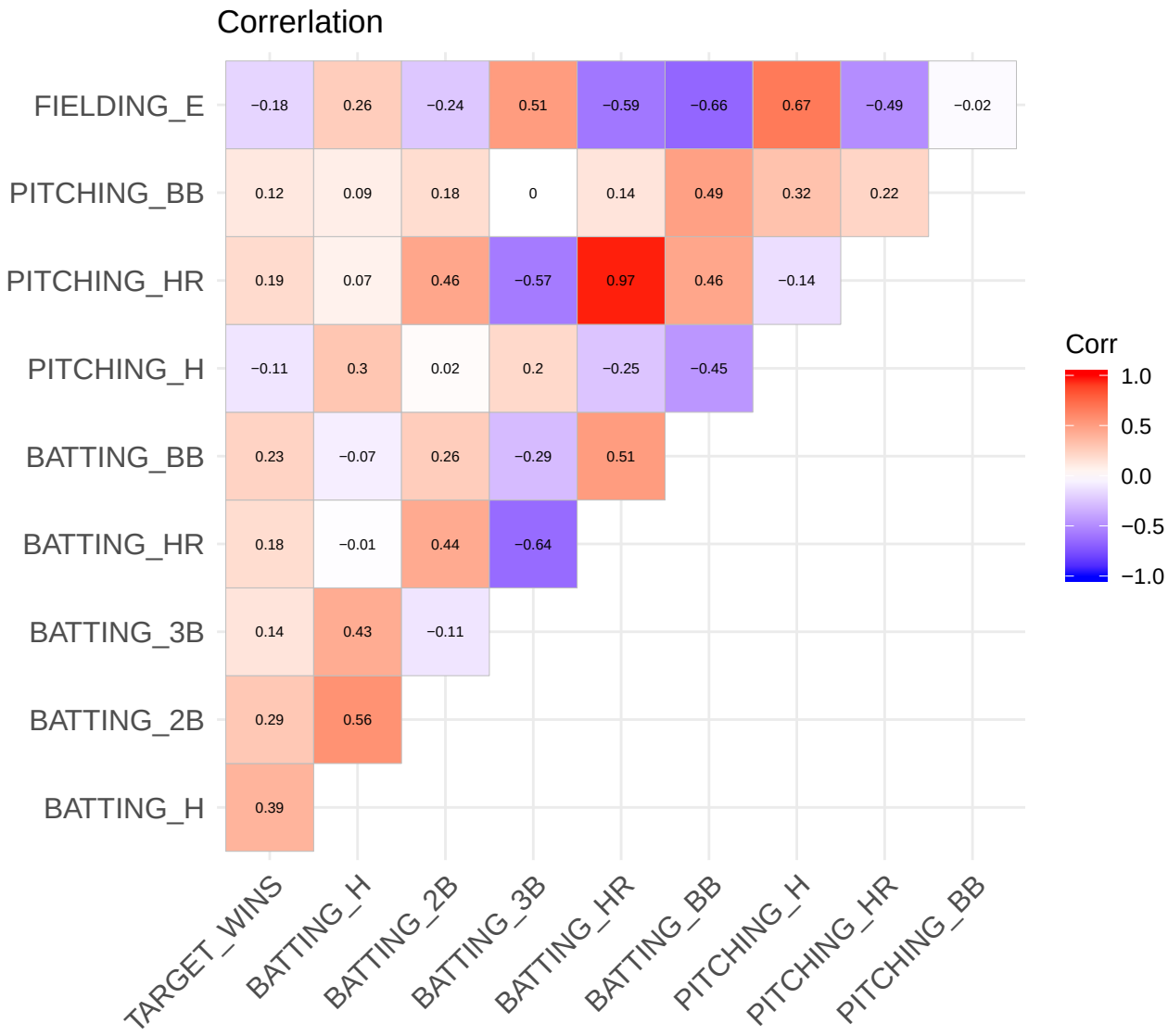


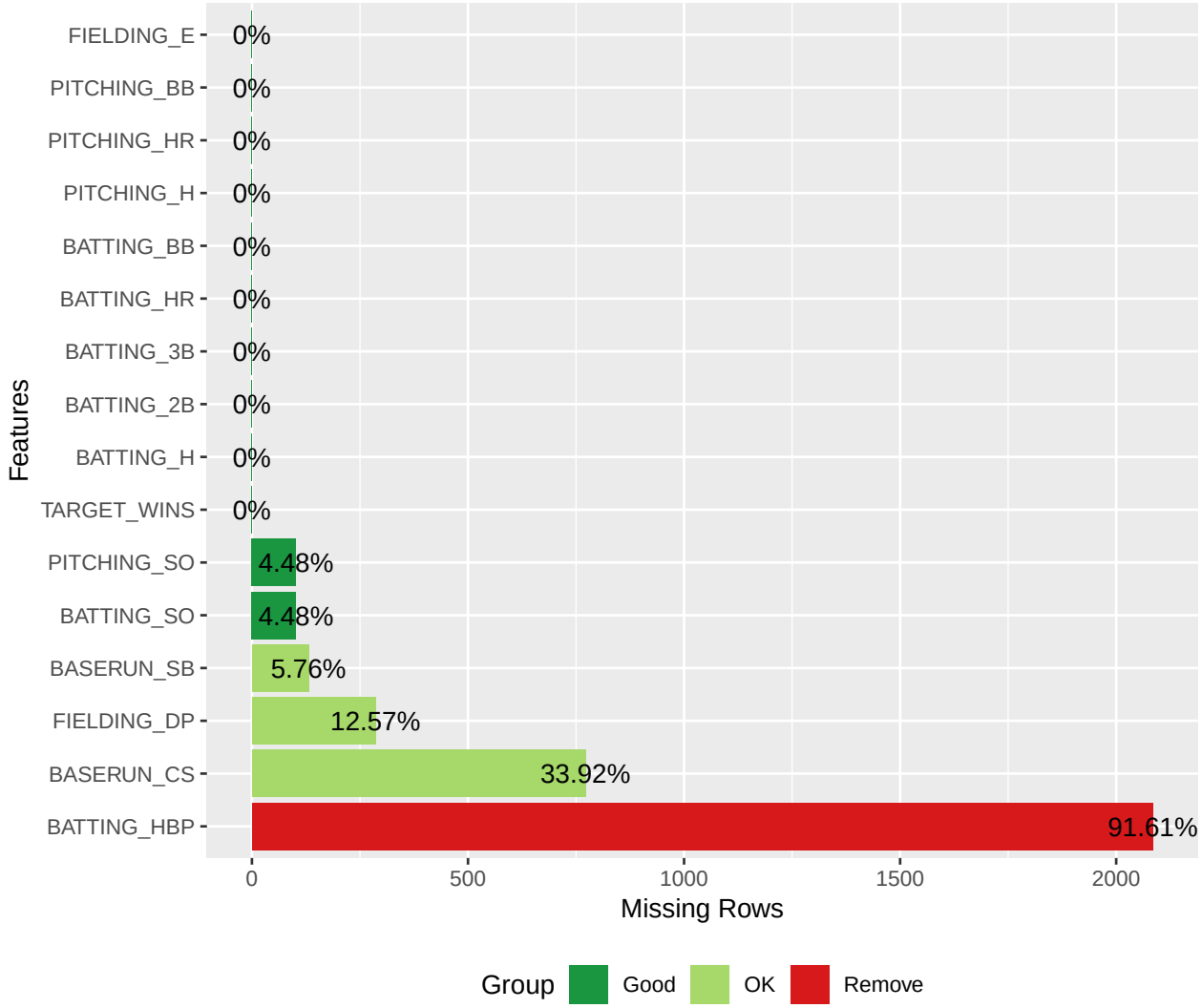
- Data visualization


Boxplot
FIELDING_DP
FIELDING_E
PITCHING_SO
PITCHING_BB
PITCHING_HR
PITCHING_H
BATTING_HBP
BASERUN_CS
BASERUN_SB
BATTING_SO
BATTING_BB
BATTING_HR
BATTING_3B
BATTING_2B
BATTING_H
TARGET_WINS
10
100
1000
10000
log transformed freq.

Scatterplot




Correrlation
FIELDING_E
PITCHING_BB
PITCHING_HR
PITCHING_H
BATTING_BB
BATTING_HR
BATTING_3B
BATTING_2B
BATTING_H
−0.18 0.26 −0.24 0.51 −0.59 −0.66 0.67 −0.49 −0.02
0.12 0.09 0.18 0 0.14 0.49 0.32 0.22
0.19 0.07 0.46 −0.57 0.97 0.46 −0.14
−0.11 0.3 0.02 0.2 −0.25 −0.45
0.23 −0.07 0.26 −0.29 0.51
0.18 −0.01 0.44 −0.64
0.14 0.43 −0.11
0.29 0.56
0.39
TARGET_WINS
BATTING_H
BATTING_2B
BATTING_3B
BATTING_HR
BATTING_BB
PITCHING_H
PITCHING_HR
PITCHING_BB
Corr
1.0
0.5
0.0
−0.5
−1.0



2 Data preparation

2.1 Possible writeup for Data Preparation

As previously mentioned, just north of 10% of the data was missing values. Missing values can lead to errors in a model, bias, and worse if left unaccounted for. Attempting to "fix" this by imputing values or guessing why the values are missing in the first place - such as concluding that the missing values are meant to be zeroes - are just as likely to help with creating a model as it is to help with creating a disaster.

One of the R packages utilized, DataExplorer, recommends removing null or missing values; it was for this reason all observations of hits by pitch were removed.<< SOURCE FOR THIS IS REQUIRED! IF NO SOURCE IS PROVIDED, CONSIDER USING "Due to the sheer volume of missing values present in the observations for hits by pitch (91.61%) it was determined the best course of action was to remove the variable altogether." OR A VARIATION THEREOF. >> Deleting all cases with missing values, in this instance, would have shrunk the size of the dataset down to less than a tenth of its original size. For this reason, the feature itself was excluded from the dataset, rather than the cases that had no values present for it.

The other missing values - present in batting strikeouts... needs more work. x_x

2.2 Information in general, don't use the writeups here as-is

2.3 Missing Values

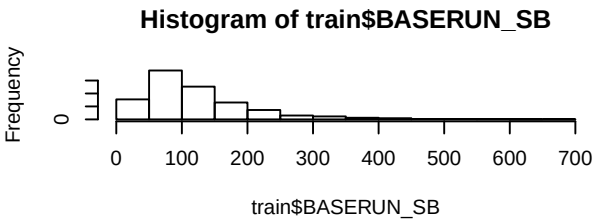
1) `Hit by pitch` missing 91.61% .

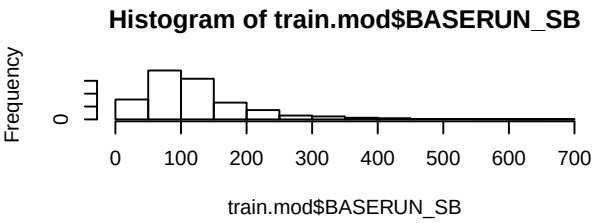
- Missing values can lead to errors and bias into a model. Fixing and imputation may help or make it worse.
- When it is just a few observations missing, modifications can be made, however, with 91.61% is a large proportion and could distort the modelling later on that it is better to ignore this column.
- The Data explorer package recommends to remove.
- From LMR: Missing Completely at Random (MCAR) The probability that a value is missing is the same for all cases. If we simply delete all cases with missing values from the analysis, we will cause no bias, although we may lose some information.
- However, there is no consensus on when to exclude missing data. Some argue that missing data more than 10% can lead to bias. Others argue that missing data patterns have greater impact than the proportion.

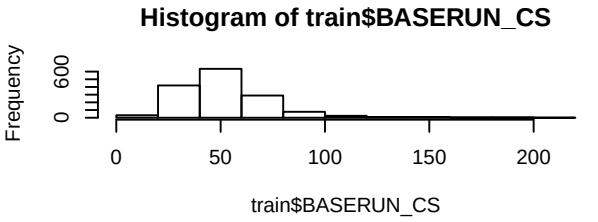
2) `Pitching_SO` and `Batting_SO` are missing exact same proportion 4.48% and are missing in the same observations.

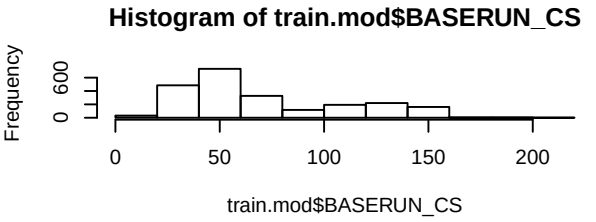
2.4 NA Imputation

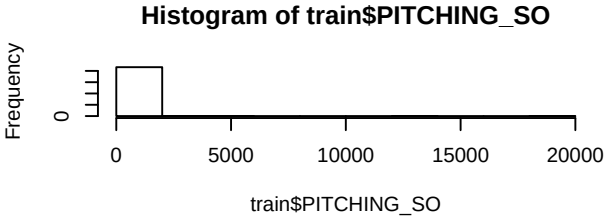
```
## TARGET_WINS   BATTING_H  BATTING_2B  BATTING_3B  BATTING_HR  BATTING_BB
##           0           0           0           0           0           0
##  BATTING_SO  BASERUN_SB  BASERUN_CS  PITCHING_H PITCHING_HR PITCHING_BB
##         102         131         772           0           0           0
## PITCHING_SO  FIELDING_E FIELDING_DP
##         102           0         286
```

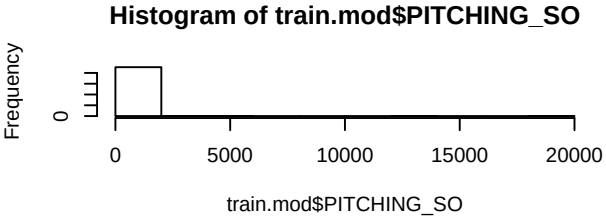


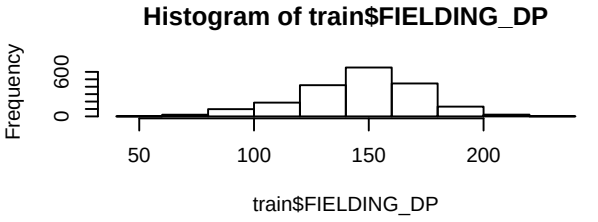


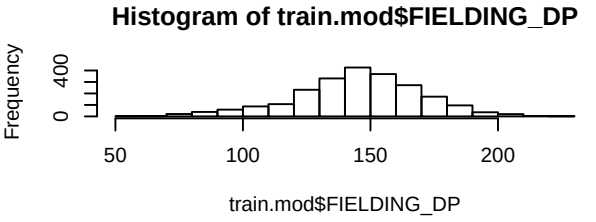












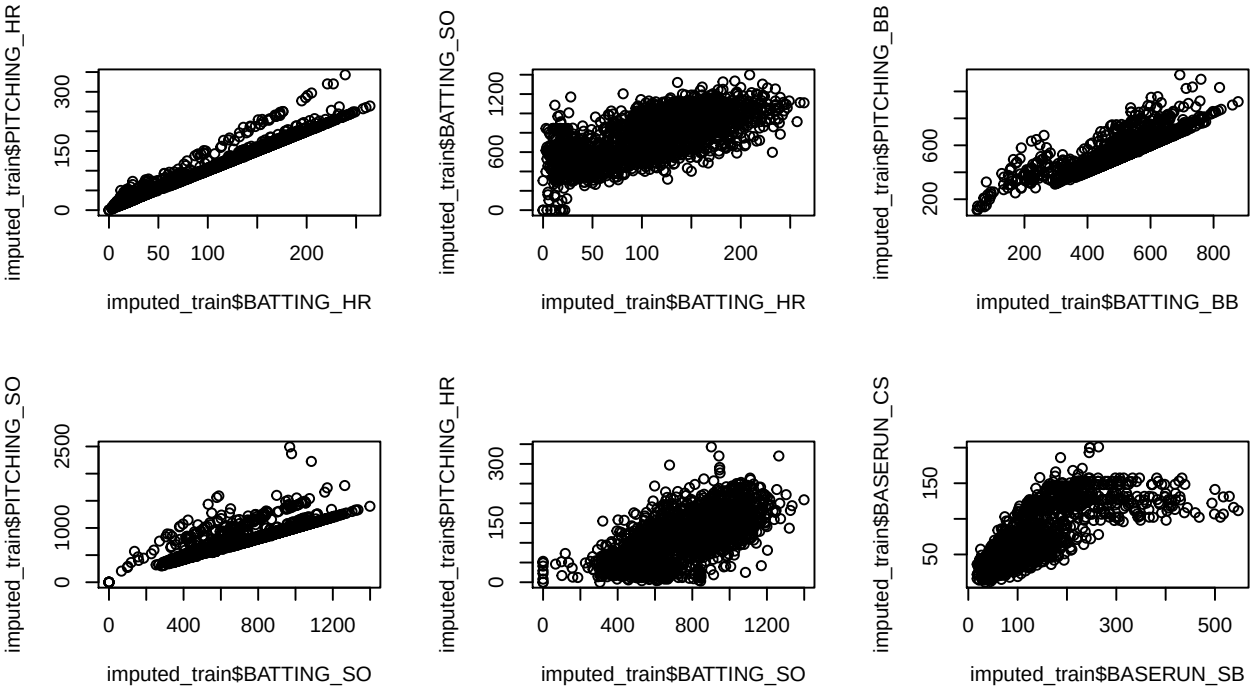
2.5 Feature Engineering

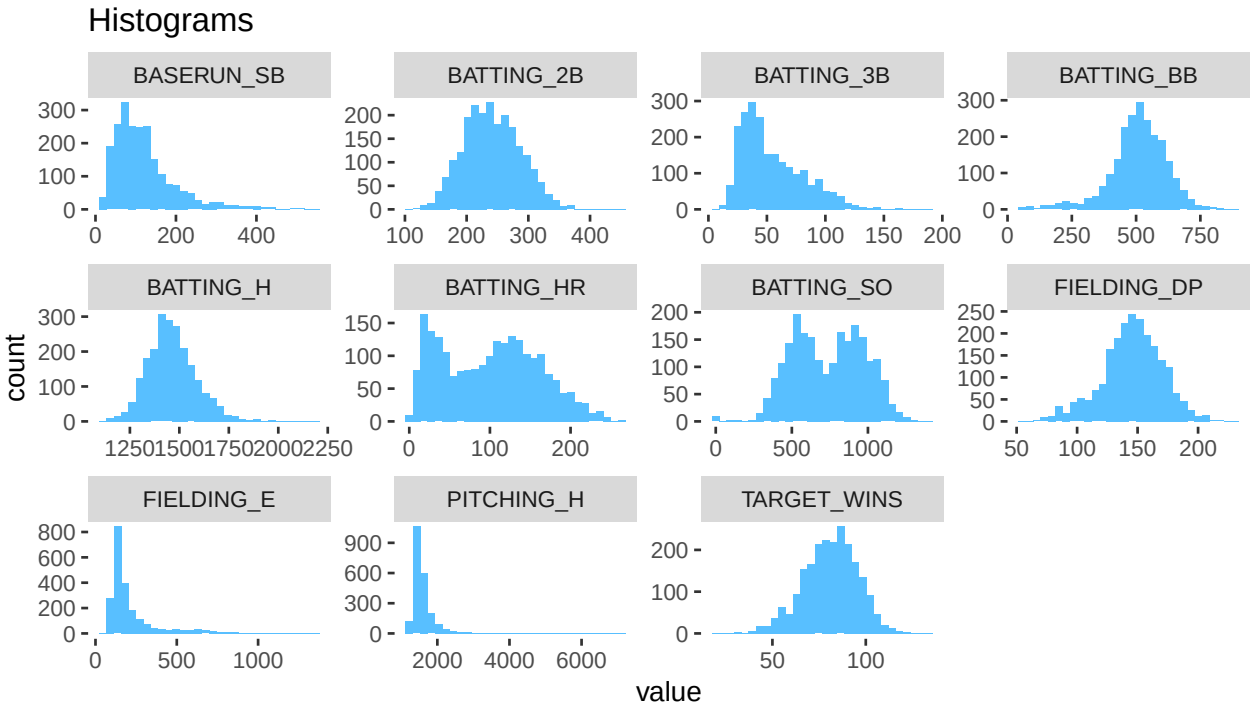
FOR THE OTHER HALF OF THE GROUP:

z_train <- sapply(imputed_train, scale) log_train <- log(imputed_train) # weird results z_log_train <- sapply(log_train, scale) # weirder results

imputed_train is most likely the variable you want to use.

3 3. Betsy's MODELS





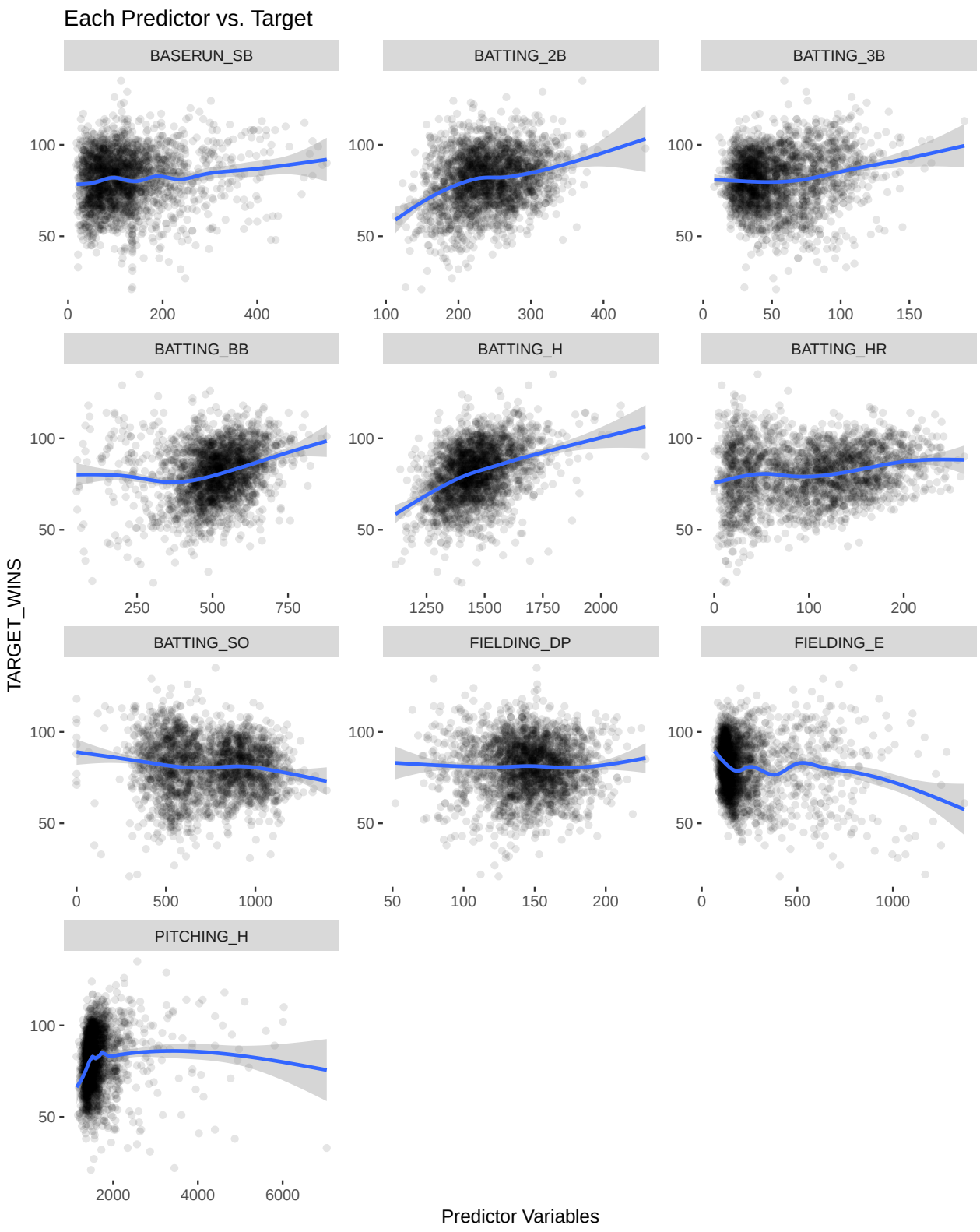
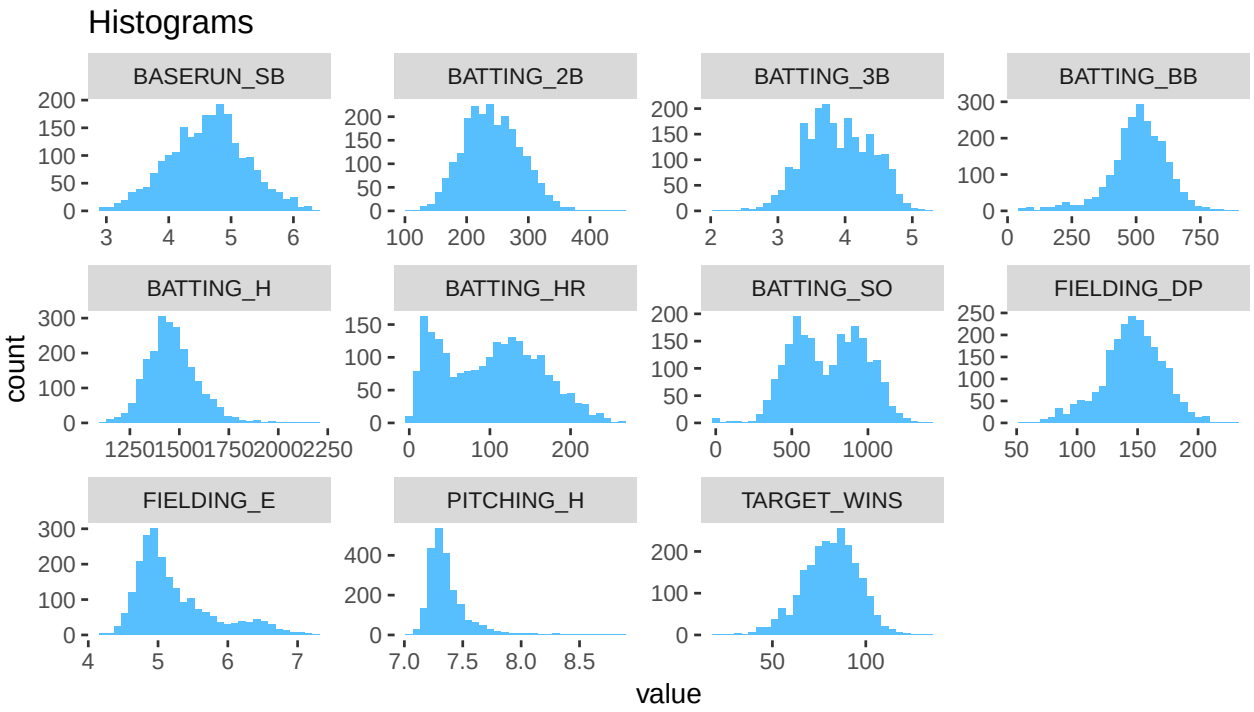


Figure 1: Each Predictor vs. Target

Table 1: Full Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-62.9724524	15.9139042	-3.957071	0.0000783
BATTING_H	0.0255615	0.0046111	5.543426	0.0000000
BATTING_2B	-0.0297598	0.0092155	-3.229309	0.0012590
BATTING_3B	6.9697812	0.9476551	7.354765	0.0000000
BATTING_HR	0.0703282	0.0101256	6.945579	0.0000000
BATTING_BB	0.0192290	0.0031872	6.033192	0.0000000
BATTING_SO	-0.0112281	0.0024113	-4.656450	0.0000034
BASERUN_SB	4.5460080	0.5617657	8.092356	0.0000000
PITCHING_H	19.1408576	2.6161541	7.316411	0.0000000
FIELDING_E	-13.1027616	1.0618606	-12.339437	0.0000000
FIELDING_DP	-0.1067225	0.0131384	-8.122924	0.0000000




```
## [1] 0.2888829
```

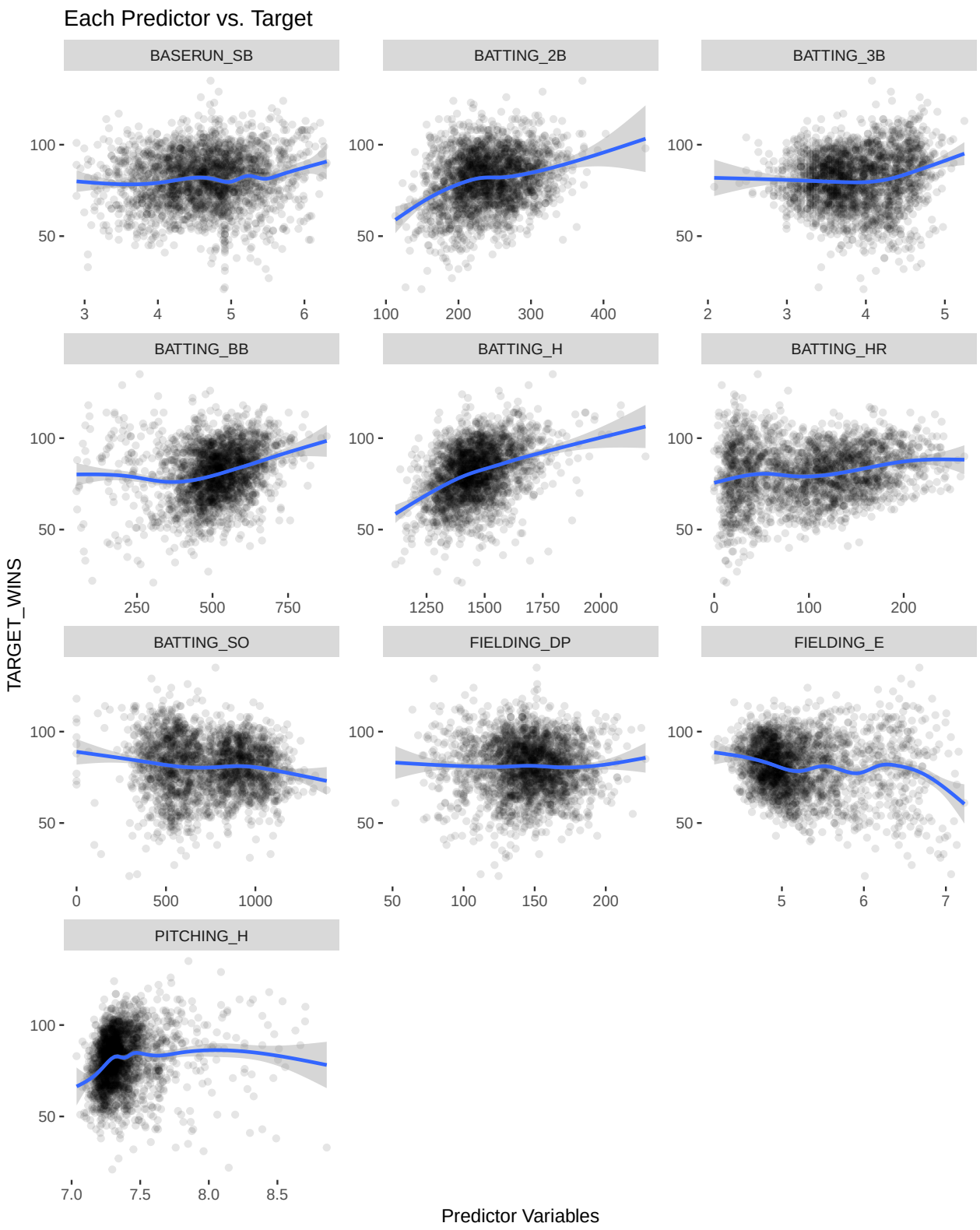
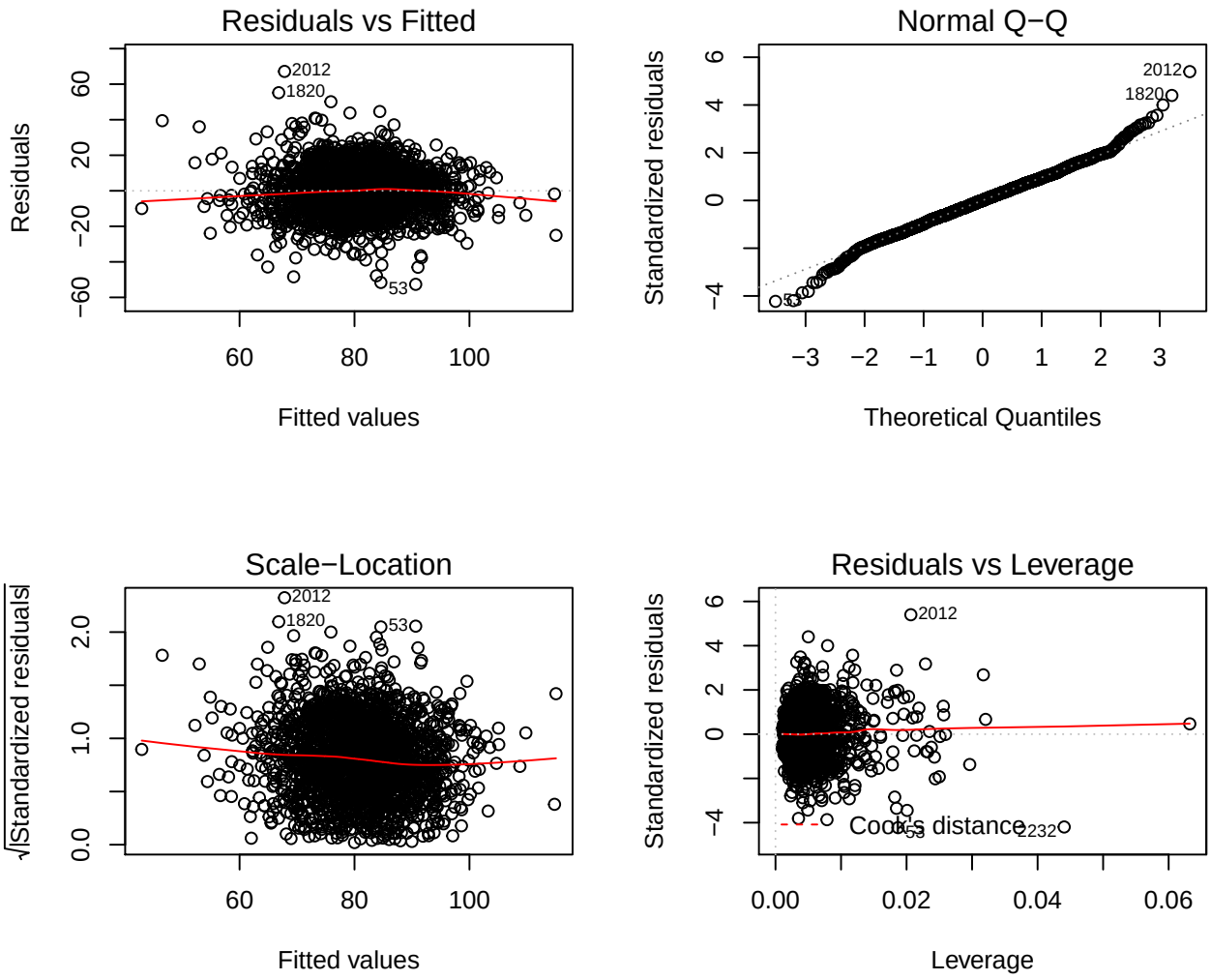
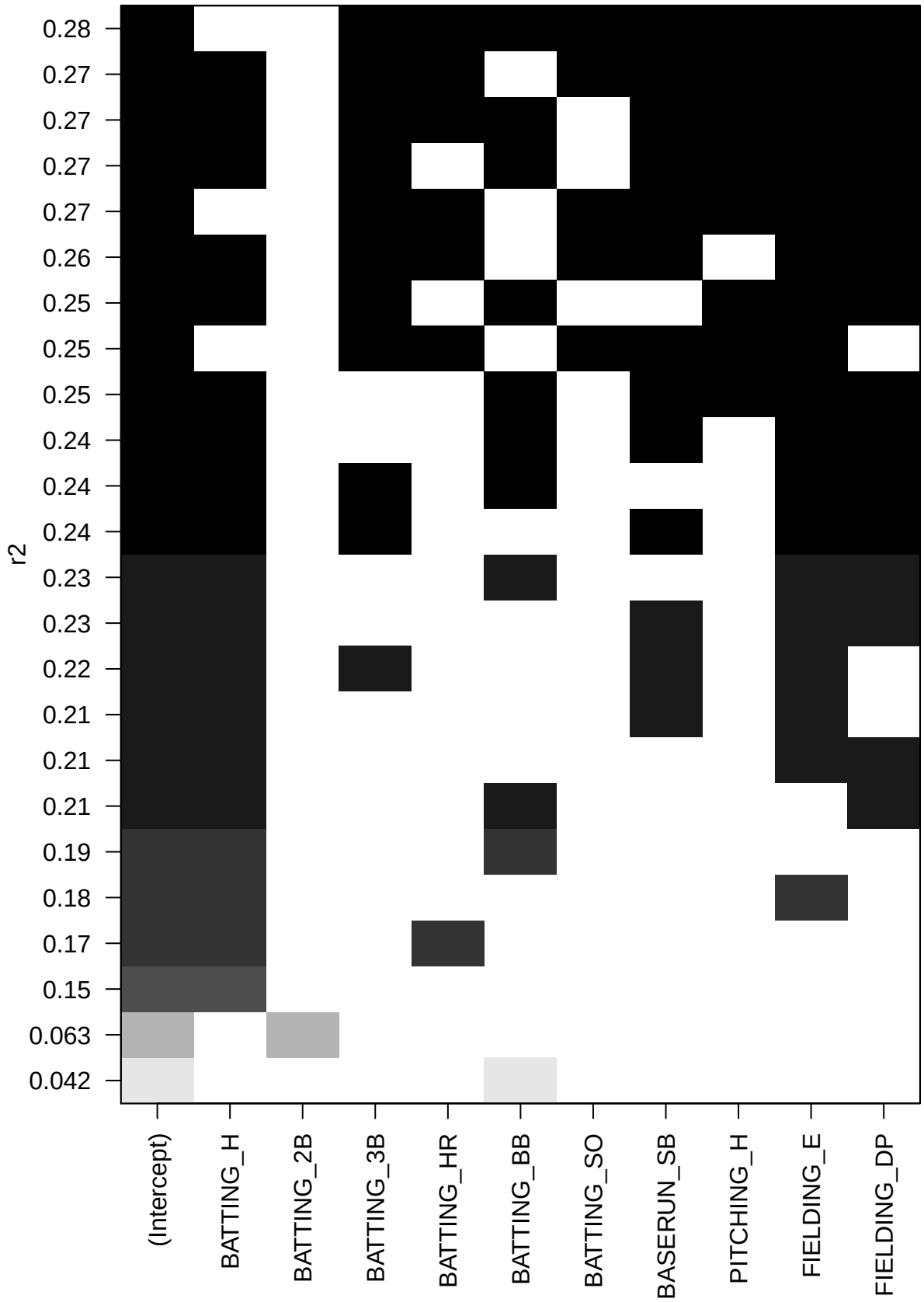


Figure 2: Each Predictor vs. Target


Residuals vs Fitted
Residuals
Fitted values
2012
1820
53
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
Leverage
232

```
## NULL
```

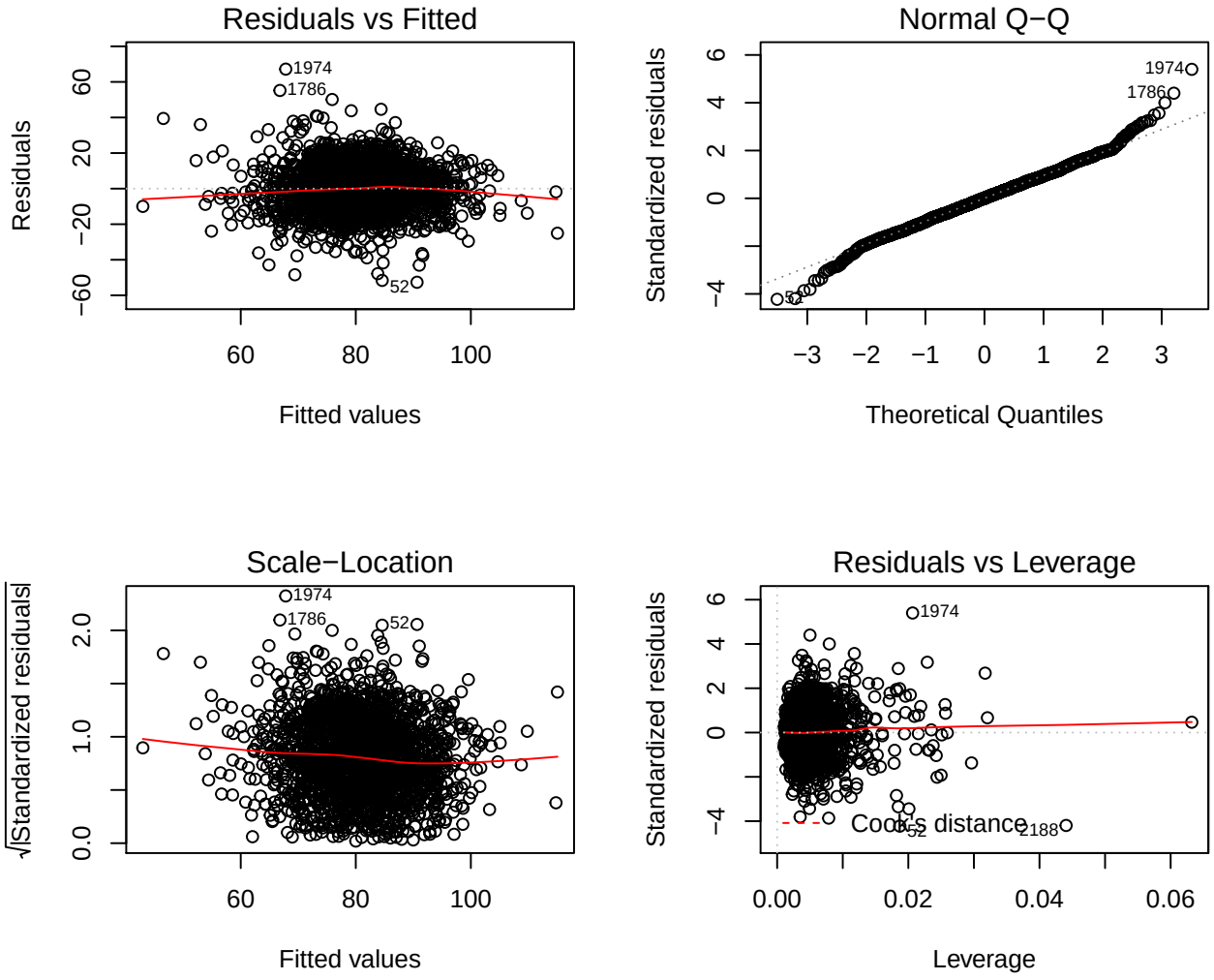

r2
0.28
0.27
0.27
0.27
0.27
0.26
0.25
0.25
0.25
0.24
0.24
0.24
0.23
0.23
0.22
0.21
0.21
0.21
0.19
0.18
0.17
0.15
0.063
0.042
(Intercept)
BATTING_H
BATTING_2B
BATTING_3B
BATTING_HR
BATTING_BB
BATTING_SO
BASERUN_SB
PITCHING_H
FIELDING_E
FIELDING_DP

```
## NULL
```

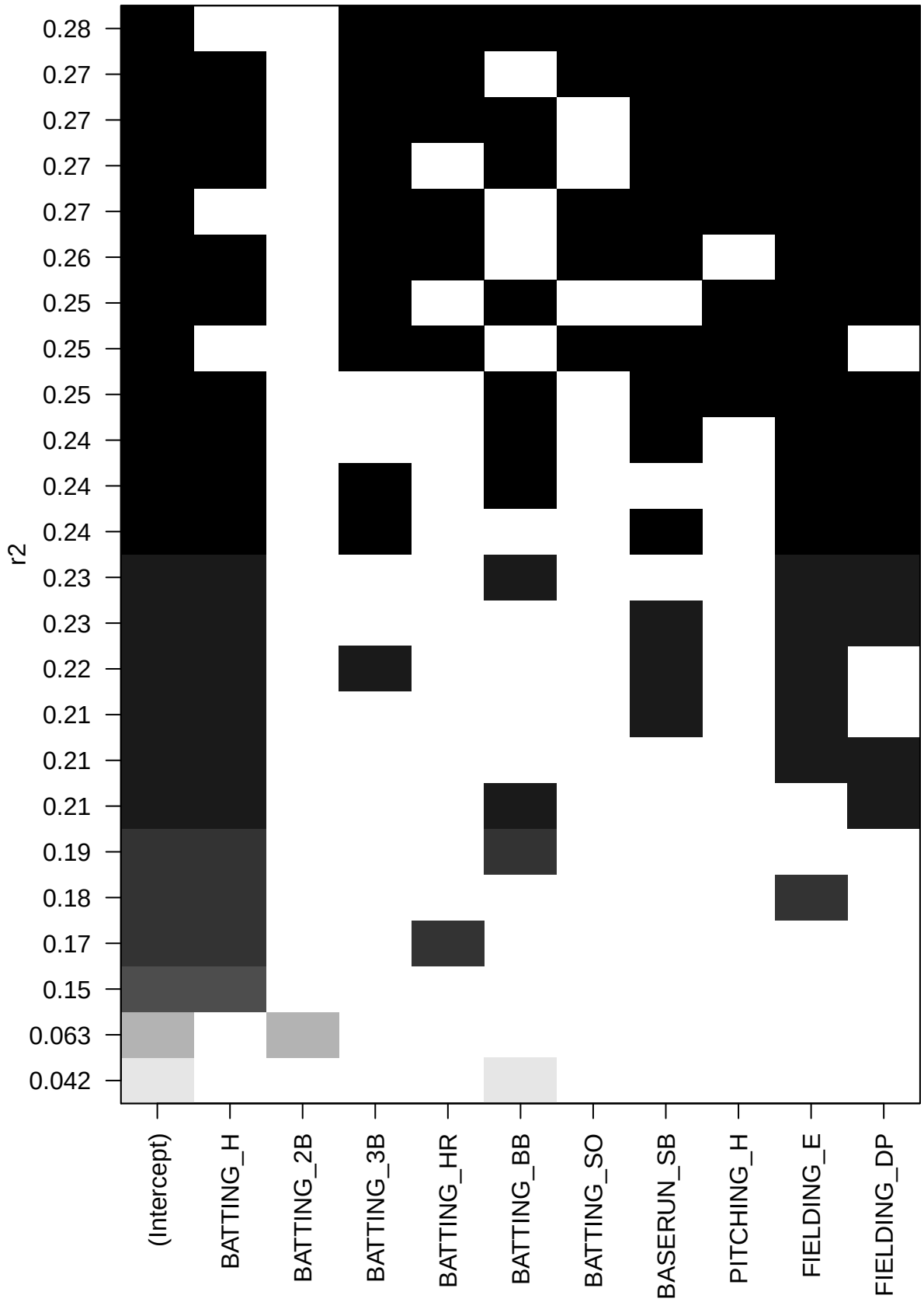
Table 2: Full SCALED Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	80.971749	0.2662860	304.078090	0.0000000
BATTING_H	3.222999	0.5814092	5.543426	0.0000000
BATTING_2B	-1.364857	0.4226469	-3.229309	0.0012590
BATTING_3B	3.381556	0.4597775	7.354765	0.0000000
BATTING_HR	4.220294	0.6076231	6.945579	0.0000000
BATTING_BB	2.190025	0.3629961	6.033192	0.0000000
BATTING_SO	-2.641712	0.5673232	-4.656450	0.0000034
BASERUN_SB	2.794642	0.3453435	8.092356	0.0000000
PITCHING_H	3.870518	0.5290186	7.316411	0.0000000
FIELDING_E	-7.441986	0.6031058	-12.339437	0.0000000
FIELDING_DP	-2.676186	0.3294610	-8.122924	0.0000000

```
## [1] 0.2888829
```




```
## NULL
```


r2
0.28
0.27
0.27
0.27
0.27
0.26
0.25
0.25
0.25
0.24
0.24
0.24
0.23
0.23
0.22
0.21
0.21
0.21
0.19
0.18
0.17
0.15
0.063
0.042
(Intercept)
BATTING_H
BATTING_2B
BATTING_3B
BATTING_HR
BATTING_BB
BATTING_SO
BASERUN_SB
PITCHING_H
FIELDING_E
FIELDING_DP

```
## NULL
```

3.0.1 Test all of the predictors

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2229	493421.2	NA	NA	NA	NA
2219	350880.3	10	142541	90.14425	0

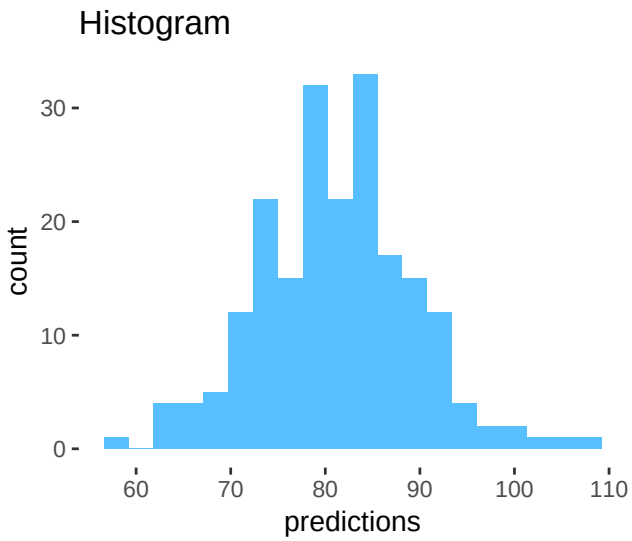
3.0.2 Test one predictor

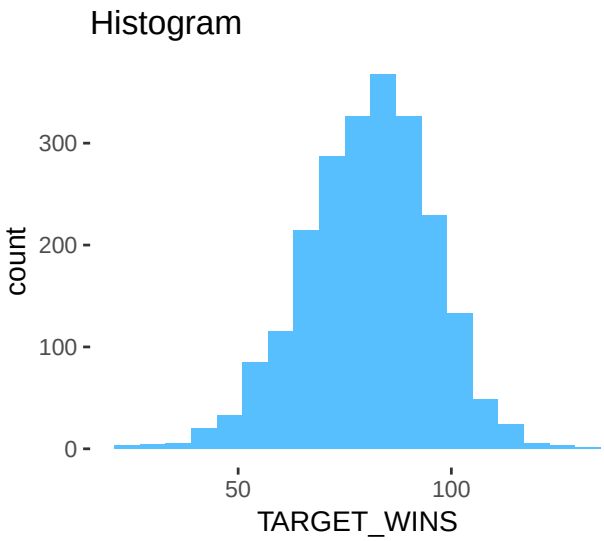
Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2220	355739.4	NA	NA	NA	NA
2219	350880.3	1	4859.127	30.72958	0

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2220	352529.3	NA	NA	NA	NA
2219	350880.3	1	1649.001	10.42844	0.001259

3.0.3 Testing a subspace

```
##
## Call:
## lm(formula = TARGET_WINS ~ I(BATTING_HR + PITCHING_HR) + I(BATTING_BB +
##     PITCHING_BB) + I(BATTING_SO + PITCHING_SO) + BATTING_H +
##     BATTING_2B + BATTING_3B + BASERUN_SB + BASERUN_CS + PITCHING_H +
##     FIELDING_E + FIELDING_DP, data = all_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.298  -8.254   0.184   8.251  66.950
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 33.769875   5.912262   5.712 1.27e-08 ***
## I(BATTING_HR + PITCHING_HR)  0.042078   0.004928   8.538  < 2e-16 ***
## I(BATTING_BB + PITCHING_BB)  0.006255   0.001547   4.044 5.44e-05 ***
## I(BATTING_SO + PITCHING_SO) -0.003880   0.001061  -3.656 0.000262 ***
## BATTING_H                    0.023533   0.004888   4.815 1.57e-06 ***
## BATTING_2B                  -0.007472   0.009482  -0.788 0.430776
## BATTING_3B                   0.104289   0.017601   5.925 3.61e-09 ***
## BASERUN_SB                   0.023849   0.005918   4.030 5.76e-05 ***
## BASERUN_CS                   0.037198   0.016067   2.315 0.020690 *
## PITCHING_H                   0.008734   0.001205   7.248 5.80e-13 ***
## FIELDING_E                  -0.035162   0.003271 -10.748  < 2e-16 ***
## FIELDING_DP                 -0.084532   0.013704  -6.168 8.18e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.74 on 2218 degrees of freedom
## Multiple R-squared:  0.2704, Adjusted R-squared:  0.2668
## F-statistic: 74.74 on 11 and 2218 DF,  p-value: < 2.2e-16
```






```
## Start:  AIC=11321.3
## TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B + BATTING_HR +
##     BATTING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H +
##     PITCHING_HR + PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP +
##     PB_H + PB_HR + PB_BB + PB_SO
##
##
## Step:  AIC=11321.3
## TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B + BATTING_HR +
##     BATTING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H +
##     PITCHING_HR + PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP +
##     PB_H + PB_HR + PB_BB
##
##
## Step:  AIC=11321.3
## TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B + BATTING_HR +
##     BATTING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H +
##     PITCHING_HR + PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP +
##     PB_H + PB_HR
##
##
## Step:  AIC=11321.3
## TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B + BATTING_HR +
##     BATTING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H +
##     PITCHING_HR + PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP +
##     PB_H
##
##
## Step:  AIC=11321.3
## TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B + BATTING_HR +
##     BATTING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H +
##     PITCHING_HR + PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP
##
##                Df Sum of Sq    RSS   AIC
## - PITCHING_HR   1      71.3 352682 11320
## - BATTING_2B    1     147.1 352757 11320
## <none>                     352610 11321
```

```
## - BASERUN_CS   1     905.6 353516 11325
## - BATTING_HR   1     932.0 353542 11325
## - BATTING_H    1    1774.0 354384 11330
## - PITCHING_SO  1    1884.8 354495 11331
## - PITCHING_BB  1    2313.2 354924 11334
## - BATTING_SO   1    3164.7 355775 11339
## - BATTING_BB   1    3725.7 356336 11343
## - BASERUN_SB   1    4387.3 356998 11347
## - FIELDING_DP  1    5849.4 358460 11356
## - BATTING_3B   1    7439.0 360049 11366
## - PITCHING_H   1   12791.3 365402 11399
## - FIELDING_E   1   20259.4 372870 11444
##
## Step:  AIC=11319.75
## TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B + BATTING_HR +
##     BATTING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H +
##     PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP
##
##               Df Sum of Sq    RSS   AIC
## - BATTING_2B   1     144.5 352826 11319
## <none>                     352682 11320
## + PITCHING_HR  1      71.3 352610 11321
## + PB_HR        1      71.3 352610 11321
## - BASERUN_CS   1     923.1 353605 11324
## - BATTING_H    1    1786.1 354468 11329
## - PITCHING_SO  1    1855.1 354537 11329
## - BATTING_SO   1    3113.1 355795 11337
## - PITCHING_BB  1    4312.8 356995 11345
## - BASERUN_SB   1    4384.3 357066 11345
## - FIELDING_DP  1    5907.2 358589 11355
## - BATTING_BB   1    6344.8 359026 11358
## - BATTING_3B   1    7380.9 360063 11364
## - BATTING_HR   1   12111.1 364793 11393
## - PITCHING_H   1   12883.5 365565 11398
## - FIELDING_E   1   21170.8 373853 11448
##
## Step:  AIC=11318.66
## TARGET_WINS ~ BATTING_H + BATTING_3B + BATTING_HR + BATTING_BB +
##     BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H + PITCHING_BB +
##     PITCHING_SO + FIELDING_E + FIELDING_DP
##
##               Df Sum of Sq    RSS   AIC
## <none>                     352826 11319
## + BATTING_2B   1     144.5 352682 11320
## + PITCHING_HR  1      68.8 352757 11320
## + PB_HR        1      68.8 352757 11320
## - BASERUN_CS   1     997.3 353824 11323
## - PITCHING_SO  1    1748.3 354575 11328
## - BATTING_H    1    1953.4 354780 11329
## - BATTING_SO   1    3092.7 355919 11336
## - PITCHING_BB  1    4257.3 357084 11343
## - BASERUN_SB   1    4451.1 357277 11345
## - FIELDING_DP  1    5838.4 358665 11353
## - BATTING_BB   1    6265.6 359092 11356
```

```
## - BATTING_3B   1    7722.8 360549 11365
## - BATTING_HR   1   12449.8 365276 11394
## - PITCHING_H   1   13383.6 366210 11400
## - FIELDING_E   1   21030.9 373857 11446
```

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	2215	352610.4	11321.30
- PB_SO	0	0.00000	2215	352610.4	11321.30
- PB_BB	0	0.00000	2215	352610.4	11321.30
- PB_HR	0	0.00000	2215	352610.4	11321.30
- PB_H	0	0.00000	2215	352610.4	11321.30
- PITCHING_HR	1	71.32528	2216	352681.7	11319.75
- BATTING_2B	1	144.53570	2217	352826.2	11318.66

```
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_H + BATTING_3B + BATTING_HR +
##     BATTING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H +
##     PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP, data = imputed_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.887  -8.130   0.041   8.097  68.359
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.175729   5.656513   5.865 5.16e-09 ***
## BATTING_H    0.013834   0.003949   3.503 0.000468 ***
## BATTING_3B   0.125448   0.018008   6.966 4.28e-12 ***
## BATTING_HR   0.090183   0.010196   8.845  < 2e-16 ***
## BATTING_BB   0.083849   0.013363   6.275 4.20e-10 ***
## BATTING_SO  -0.029123   0.006606  -4.408 1.09e-05 ***
## BASERUN_SB   0.032719   0.006187   5.289 1.35e-07 ***
## BASERUN_CS   0.040007   0.015982   2.503 0.012375 *
## PITCHING_H   0.014048   0.001532   9.170  < 2e-16 ***
## PITCHING_BB -0.060007   0.011602  -5.172 2.52e-07 ***
## PITCHING_SO  0.017834   0.005381   3.314 0.000933 ***
## FIELDING_E  -0.037924   0.003299 -11.496  < 2e-16 ***
## FIELDING_DP -0.082243   0.013579  -6.057 1.63e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.62 on 2217 degrees of freedom
## Multiple R-squared:  0.2849, Adjusted R-squared:  0.2811
## F-statistic: 73.62 on 12 and 2217 DF,  p-value: < 2.2e-16
```